On Recognition of Declarative Questions in English

Marie Šafářová & Marc Swerts

ILLC, University of Amsterdam and
Tilburg University, The Netherlands
M.Safarova@uva.nl, m.g.j.swerts@uvt.nl

Abstract

We report on the results of an experiment designed to test the phonological properties of declarative questions in American English. Previous work is not conclusive in whether a typical melodic contour exists for declarative questions and whether it is crucial for their recognition in spontaneous dialogues. We conclude that speakers are equally good at recognizing declarative questions with and without having access to prosodic information. However, certain contours are taken to be more likely to signal questions, especially the rising contour described in Gunlogson (2001).[1]

1. Introduction

Traditionally, it has been assumed that declarative questions (DQs) in American English are recognized by their intonational properties (e.g., 'rising contour', 'high boundary'). Pierrehumbert & Hirschberg (1990) consider L* H H% to be the typical question contour and we could thus expect it to be the phonological realization of the DQ contour as well. Bartels (1999) suggests that not only L* H H%, but also L* H L%, H* H H% and H* H L% are "non-assertive contours" in English; in the terminology of Bartels & Merin (1997), they alienate choice over the status of the expressed proposition to the addressee. Gunlogson (2001), assumes that one of the necessary properties of declarative questions is their rising intonation, which she defines as "non-falling from the nuclear pitch accent to the terminus and ending at a point higher than the level of the nuclear accent", a description which (in her own understanding) fits all of the tunes H* H H%, L* H H%, L* H L% and L* L H%. Adopting Gunlogson's approach in the relevant aspects, Steedman (2003) considers boundary tones to be crucial with respect to speaker's or hearer's commitment, suggesting that H H% and L H% express hearer's commitment (and are thus presumably associated with declarative questions) while L L% and H L% express speaker's commitment. To summarize, there seems to rule a consensus with respect to the assumption that declarative questions are marked with 'question intonation' but the exact nature of the contour is a subject of disagreement.

[1]The authors would like to thank Gilad Mishne, Maarten de Rijke and Juan Heguiabehere from the Inference Technology Group at the ILLC who kindly provided technical help with the experimental setup, and Brian MacWhinney for making the Santa Barbara corpus, part II, available. Special thanks to Stefan Benus, Laurie Maynell and Julie McGory for annotations, as well as to Jocelyn Ballantyne, Paul Dekker, Pieter Koele, Ivana Kruijff-Korbayová and Craige Roberts for their comment. We are grateful to the numerous people who either participated in the experiments themselves or helped to look for suitable subjects. The experiments were also presented at www.linguistic-experiments.org.
The authors are solely responsible for any errors and/or misconceptions.

How necessary is intonation for recognition of declarative questions? As for yes-no questions, it has been shown that they do not always occur with a rising intonation in natural data[2] (Hirschberg (2000) found that about 30% in read speech and 43% in spontaneous speech were falling). One could argue that intonational marking is not necessary for yes-no questions, given that they are already marked by syntactic inversion. In fact, Haan (2001) formulates this assumption in terms of her *Functional Hypothesis*, which predicts that high question pitch will be maximally present in questions that are not otherwise marked for interrogativity (i.e, declarative questions), somewhat less in questions with inversion, and least in questions with both a question word and inversion. Her experimental research confirms the hypothesis for Dutch: all the declarative questions in her corpus were rendered with a rising pitch, albeit her study is based on read speech provided with clear punctuation, such as question marks. On the other hand, there is evidence that the question status of utterances with declarative syntax may be cued by non-intonational features as well. For instance, Beun (also for Dutch) found that in his corpus of natural dialogues, about 20% of declarative questions were falling. He also noted that, although declarative questions were much less likely to be used in written conversations, they could often be correctly identified if they contained second person personal pronouns, an expression of uncertainty and/or particles like *en* ('and'), *dus* ('so') or *ook* ('also') at the beginning of the utterance (Beun 1989, 1990).

Beun's results are in line with the outcome of an experimental study by Geluykens (1987) who found some utterances with a declarative syntax to be more 'question-prone' than others. For instance, a sentence like "You feel ill" is likely to have an interrogative intent, since one cannot easily make a statement about the inaccessible internal state of another person. On the contrary, "I feel ill" is more statement-prone, as the speaker is not likely to question his/her own feelings. Using such sentences provided with artificial rising and falling contours, Geluykens found that the relative cue value of rising intonation as a marker of questions very much depended on lexical-pragmatic properties of utterances. In follow-up studies using spontaneous speech corpora of Southern British English, Geluykens (1988) found that a majority of declarative questions in his corpus occurred with a fall (57% of the data, with the overall frequency of falls - 64%). On the basis of his research, he concluded that intonation is "virtually irrelevant as a question cue" (Geluykens 1988:479) and that lexical-pragmatic indicators are more important for determining the question status of an utterance.

The discussion of the literature thus brings to light that

[2]For wh-questions, it has always been assumed that they were mostly falling.

there is some doubt as to the intonational properties of DQs, and their importance in comparison to lexical-pragmatic markers of questionhood. However, past studies have had some important methodological drawbacks. On the one hand, in a speaker-oriented corpus study, it is often difficult to adequately operationalize what exactly constitutes a DQ; researchers have usually based their classifications on the addressee's reaction, which gives only circumstantial evidence about the pragmatic status of the prior utterance. On the other hand, more controlled experimental approaches, using read or synthetic speech, have been criticized for using stimulus utterances with intonational properties which may not be representative of natural language behavior. Therefore, the current study aims to combine techniques from these two traditions of research. First, following methods outlined by Haan and Geluykens, we will approach the problem by largely taking a listener perspective: while it is difficult to fully prove whether or not a speaker had originally intended his/her sentence to function as a question, it is more feasible to ask subjects to judge a set of utterances regarding their pragmatic status, both in speech-only and transcription-only tests. Second, our stimuli for the perception studies will consist entirely of samples taken from spontaneous speech corpora, which gives us a better guarantee that the contours in our data are "real".

2. Experimental approach

2.1. Design

In order to find out how speakers recognize declarative questions in American English and what contours are prevalently associated with them, we carried out an experiment with two parts: a non-acoustic recognition task and an acoustic recognition task. In the non-acoustic experiment, subjects were presented only with a transcription of the stimuli (with no initial capitals and no punctuation). In the acoustic experiment, another group of subjects was presented with a sound recording of the stimuli without transcriptions.

2.2. Stimuli

We made use of natural data selected from the spontaneous conversations in the Santa Barbara Corpus of Spoken American English (part I and II). From the corpus, we selected mostly mono-clausal sentences with no ellipsis and clearly indicative syntax (no subject-finite verb inversion, no wh-words). In order to obtain a stimulus set representative of the overall distribution of declarative sentences in our corpus, the utterances came from three contexts and were labelled as either (i) declarative question, (ii) acknowledged declarative, or (iii) declarative proper. The classification into these three categories was based on the type of reaction from the addressee in the subsequent context. As declarative questions we selected utterances that were turn-completing and in the context could be turned into a polar question followed by a reply that (contextually) entailed a *yes/no/I don't know* answer. As acknowledged declaratives, we chose utterances followed by a short acknowledgement ('backchannel') by the addressee but not turn-completing (the speaker immediately continued). For the category of declaratives proper, we selected utterances that were not turn-completing and were immediately followed by another utterance by the same speaker (receiving no *yes/no* response from the addressee). In this last category, we selected only those declaratives that were not followed by a clause starting with a sentence connective. The first selection was made on the transcripts of the corpus, without taking into consideration any prosodic information. The punctuation marks chosen by the corpus transcribers were disregarded. Since there is an overlap in the lexical expressions used for backchannels (used to acknowledge the speaker's contribution without claiming the floor, and indicative of acknowledged declaratives) and agreements used to assert an opinion and indicative of declarative questions (Shriberg et al. 1998), for some ambiguous cases, we considered also the prosodic properties of the responses to the selected utterances. There were 93 sentences in total (31 of each type) by 31 speakers (15 female, 16 male) with all three types per speaker. Note that, of the three utterance categories, those of type (i) are most likely to have been intended by the speakers as true questions, though the response they received depended on how cooperative a listener was in a particular dialogue.

2.3. Method

Thirty-four subjects (25 female, 9 male), all native speakers of American English, participated as judges in the two experiments (seventeen subjects in each). Their ages varied between 16 and 57. They received no financial retribution but they had a chance of winning a gift certificate for $50. The stimuli were presented to them with an interactive computer program (ww-stim) on the Internet. Regarding the speech version of the test, there was no strict control over the circumstances under which the experiment was performed (sound level, ambient noise, type of headphones, type of loudspeaker, etc.), it was only recommended that subjects do the experiment in a quiet environment and with the use of headphones. They were instructed to categorize each stimulus as either one of three categories, described in terms of expected responses to the utterance: 1. *speaker will continue*, 2. *addressee will show (s)he understands and speaker will continue*, 3. *speaker wants the addressee to confirm or disconfirm speaker's statement*. These three types of responses were taken to correspond to proper declaratives, acknowledged declaratives and declarative questions, respectively. In both experiments, the presentation of the stimuli was randomized in order to make up for possible learning effects. Subjects needed on average approximately 11 minutes to complete the non-acoustic experiment and 27 minutes for the acoustic experiment (due to download time of the sound files).

3. Results

3.1. Non-acoustic task

Table 1 summarizes the overall classification of the declarative types in the non-acoustic task (for each type, 31 utterances $\times$ 17 subjects). The classification (without statistic significance) for declarative questions was better than the classification of the other two categories, with a slight majority of DQs classified correctly. At first sight, this table suggests that it was very hard for subjects to estimate whether the utterances were originally followed by a full, short or no reaction from an addressee. However, if we only concentrate on the "clear" cases, i.e., ut-

Table 1: *Classification of declarative types in non-acoustic task (without statistical significance).*

Declarative Type	Correct	Incorrect
Declarative Proper	184	343
Acknowledged Declarative	212	315
Declarative Question	276	251

terances that got a statistically significant classification (based on χ^2 tests, $p < .05$), we get a somewhat different picture (see table 2). There it can be seen that judges tend to be able to distinguish the DQs from the other two utterance types (16 out of 31 DQs were correctly identified). This result for American

Table 2: *Significant classification of declarative sentences in non-acoustic task.*

Judged as: / Decl.Type	DP	AD	DQ
DP	5	4	1
AD	4	3	2
DQ	1	3	16

English corresponds to Beun's and Geluykens' conclusion for Dutch and British English, respectively, that lexical-pragmatic features play a role in question recognition. To further our understanding of this finding, we subsequently analyzed the data from a purely perceptual perspective (disregarding the original context from which the utterances were extracted), only looking at how subjects interpreted the stimulus utterances. Here, we reduced the original ternary distinction into a binary one between DQs versus nonDQs. Using Geluykens' and Beun's observations, we distinguished three binary features which could have played a role in subjects' decision process:

1. **you-presence** - presence or absence of a second person personal pronoun (in any syntactic position);
2. **I-presence** - presence or absence of a first person pronoun (in any syntactic position);
3. **particle** - presence or absence of *and/but/so/oh* at the start of the utterance.

Note that, based on previous work, we would expect opposite trends in our judgments for the presence/absence of first and second pronouns respectively, with a higher proportion of utterances classified as DQs when "you" is present, while the opposite is true for DQs containing an "I". Table 3 gives the average proportion of utterances classified as DQ as a function of the presence or absence of the three binary features, and the corresponding Mann-Whitney U stats to see whether the difference in average proportion is significant.

Table 3: *Import of lexical features to question recognition.*

Feature	Level	Av. Prop	Mann-Whitney U
You	Present (n=31)	.57	U=318
	Absent (n=62)	.25	p< 0.001
I	Present (n=23)	.29	U=700
	Absent (n=70)	.37	p=.348
Particle	Present (n=13)	.36	U=458
	Absent (n=80)	.33	p=.490

The table reveals that from the hypothesized lexical cues only the second person pronoun significantly distinguishes DQ classifications from nonDQ ones. Put differently, the second person pronoun was present on 14 of the 16 utterances significantly categorized as DQs and only on 3 of the 20 significantly categorized as nonDQs. The trend in the data for presence versus absence of first person pronoun is in the expected, opposite direction, but it is not significant significant, like the effect of the particle.

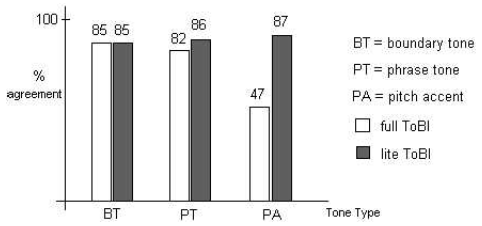

Figure 1: *Pairwise comparison of inter-annotator agreement in full ToBI and ToBI-lite.*

3.2. Acoustic task

Table 4 summarizes the overall classification of the declarative types in the acoustic task, giving a picture which is comparable to the non-acoustic results shown in table 1. Table 5 shows, as before, the categorization of declarative types with statistical significance ($p < .05$). Similarly as in the acoustic task, declarative question recognition was quite good (14 of the 31 DQs were classified as such with high significance).

Table 4: *Classification of declarative types in acoustic task (without statistical significance).*

Declarative Type	Correct	Incorrect
Declarative Proper	255	272
Acknowledged Declarative	188	339
Declarative Question	266	261

Table 5: *Significant classification of declarative sentences in acoustic task.*

Judged as: / Decl.Type	DP	AD	DQ
DP	12	1	2
AD	10	2	1
DQ	6	0	14

As in the non-acoustic test, we again analyzed the data from a purely perceptual perspective, in particular focusing on the cue value of different intonational contours. To this end, all the utterances used in the experiment were labelled with MAE-ToBI by three professional annotators. The labels were compared using pairwise agreement (viz Syrdal & McGory 2000). In order to achieve a better inter-annotator agreement, the ToBI labels were changed into a ToBI-lite version as follows: all downstepped accents were matched with their non-downstepped version and all complex pitch accents with the prominent (monotonal) pitch accent (cmp. Pitrelli et al. 1994), e.g., L+H* with H*, H+!H* with !H*.[3] While the pairwise agreement for pitch accents was low in the full ToBI version (47%), in the ToBI-lite version it was comparable to the agreement for phrase tones and boundary tones (see Figure 1). There were four binary features which were evaluated with respect to the subjects' judgements:

1. **Steedman** - an utterance would receive the value 1 if it ended with a high boundary tone and 0 if it ended with a low boundary tone;

[3] In effect, e.g., a bitonal pitch accent H+!H* would thus match with H*.

2. **Gunlogson** - with value 1 if an utterance had one of the following contours: H*H-H%, L*H-H%, L*H-L%, L*L-H%, and 0 otherwise;

3. **Bartels** - with value 1 if an utterance had one of the following contours: L*H-%H, L*H-L%, H*H-H%, H*H-L% and 0 otherwise;

4. **TQC** - with value 1 for utterances that had the "typical question contour" L*H-H% and 0 otherwise.

The features were evaluated using the ToBI-lite annotations described above. In case of disagreement, preference was given to the majority opinion.[4] For example, if an utterance was characterized as having a Gunlogson question contour by two annotators, it was assigned value 1 for this feature.

The results of the analysis are shown in table 6. The table shows that the different instantiations of question intonation can all significantly separate DQs from nonDQs, but there are distributional differences. The TQC feature appears to be the best predictor but given that there were only four agreed instances of it in the data, it is not possible to draw any conclusions about its meaning. The Gunlogson feature predicts better than the remaining two (also given that of the 17 utterances significantly classified as DQ, 12 had the feature, compared to only 2 of the utterances significantly classified as nonDQ).

Table 6: *Import of intonational features to question recognition.*

Feature	Level	Mean Rank	Mann-Whitney U
Steedman	Present (n=21)	.54	U=341
	Absent (n=72)	.22	$p < 0.001$
Gunlogson	Present (n=18)	.62	U=208.5
	Absent (n=75)	.21	$p < 0.001$
Bartels	Present (n=35)	.46	U=511
	Absent (n=58)	.18	$p < 0.001$
TQC	Present (n=4)	.81	U=19
	Absent (n=89)	.27	$p < 0.001$

4. Discussion and Conclusions

The results of the experiment show that speakers of American English are able to recognize some declarative questions whether or not they have access to prosodic information. The fact that only about a half of the DQs was classified correctly in either of the tasks suggests that contextual information plays an important role for question recognition in spontaneous dialogues. The experiment also showed that some contours are more likely to be perceived as signalling questions, in particular the ones described by Gunlogson (2000). As for lexical-pragmatic properties influencing DQ-recognition, the presence of a second person pronoun in any syntactic position in the utterance was relevant.

The two sets of utterances classified significantly as questions in the two experiments were not equal, which means that at least in some cases, prosody (and, we assume, mainly intonation) contributes decisively to question recognition. Given that in the acoustic task, subjects did not have access to transcripts and we thus cannot be sure that they understood the utterances correctly, at present it is not possible to describe these cases with reliability. It remains an interesting empirical question how subjects perceive combinations of intonational and lexical markers of questionhood. In particular, it is useful to explore whether these two types of cues support each other so that they need to co-occur, or whether one indicator of a question can overrule the cue value of another that suggests an opposite category. Finally, this research can be extended to include other possible lexical and intonational markers, and also to test the import of the larger discourse context in which a declarative utterance occurred.

[4] With respect to the presence of Steedman feature, annotators disagreed in total 23 times, for Gunlogson feature 17 times, for Bartels feature 21 times and for the Typical Question Contour (TQC) feature, 6 times, out of 93 utterances.

5. References

[1] Bartels, C., 1999. *The Intonation of English Statements and Questions. A Compositional Interpretation.* New York & London, Garland Publishing Inc.

[2] Beun, R.J., 1989. Declarative question acts: Two experiments on identification. In *The Structure of Multimodal Dialogues*, M.M. Taylor; F. Néel; D.G. Bouwhuis (eds). Amsterdam: North Holland, 313-321.

[3] Beun, R.J., 1990. The recognition of Dutch declarative questions. *Journal of Pragmatics* 14, 39-56.

[4] Geluykens, R., 1987. Intonation and speech act type. An experimental approach to rising intonation in queclaratives. *Journal of Pragmatics* 11, 483-494.

[5] Geluykens, R., 1988. On the myth of rising intonation in polar questions. *Journal of Pragmatics* 12, 467-485.

[6] Gunlogson, C., 2001. *True to Form: Rising and Falling Declaratives as Questions in English.* PhD., UCSC.

[7] Haan, J., 2001. *Speaking of Questions: An Exploration of Dutch Question Intonation*. PhD., Univ. Nijmegen.

[8] Hirschberg, J., 2000. A corpus-based approach to the study of speaking style. In *Prosody: Theory and Experiment*, M. Horne (ed.). Kluwer, 335-350.

[9] Merin, A.; Bartels, C., 1997. Decision-theoretic semantics for intonation. *Arbeitspapiere des Sonderforschungsbereichs 340*, Bericht Nr. 88. Univ. Stuttgart, Univ. Tübingen, IBM Deutschland.

[10] Pierrehumbert, J.; Hirschberg, J., 1990. The meaning of intonational contours in the interpretation of discourse. In *Intentions in communication*, P.R. Cohen; J. Morgan; M.E. Pollack (eds). Cambridge, Mass., MIT, 271-311.

[11] Pitrelli, J.; Beckman, M.; Hirschberg, J., 1994. Evaluation of prosodic transcription labeling reliability in the ToBI framework. In *Proc. ICSLP*, Yokohama, 123-126.

[12] Shriberg, E.; Bates, R.; Stolcke, A.; Taylor, P.; et al. 1998. Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and Speech* 41(3-4), 439-487.

[13] Steedman, M., 2003. Information-Structural Semantics for English Intonation. *Proc. LSA Workshop on Topic and Focus*, Santa Barbara, July 2001.

[14] Syrdal, A.K.; McGory, J., 2000. Inter-transcriber Reliability of ToBI Prosodic Labeling. In *Proc. ICSLP*, vol.3, Beijing, 235-238.